DOCUMENT RESUME

ED 090 462 CG 008 853

AUTHOR Olsen, Henry D.; Barickowski, Robert S.
TITLE Test Item Arrangement and Adaptation Level.
PUB DATE Apr 74
NOTE 15p.; Presented at the American Educational Research Association Annual Meeting (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Adaptation Level Theory; *Criterion Referenced Tests; Elementary School Students; *Item Analysis; Measurement Techniques; Perception; Research Projects; Speeches; *Test Construction
IDENTIFIERS Helsons Adaptation Level Theory

ABSTRACT

Based on a rationale provided by Helson's adaptation level theory it was predicted that students would perceive items arranged in a hard-medium-easy order as being easier than the same items arranged easy-medium-hard. The evidence presented in this study generally confirmed the preceding conjecture, but also found the Ss' perceptions did not significantly influence their scores. That is, using true-false or multiple choice items, Ss' received the same test scores regardless of the context in which they found the items. (Author)

Medgar Evers College

OF THE CITY UNIVERSITY OF NEW YORK

1127 CARROLL ST.
BROOKLYN, NY 11225

ED 090462

Test Item Arrangement and Adaptation Level

Henry Olsen, Ph.D.
Associate Professor of Teacher Education
Medgar Evers College
City University of New York
Brooklyn, New York

Robert S. Barcikowski, Ph.D.
Associate Professor of Educational Research,
Statistics and Evaluation
Ohio University
Athens, Ohio

CG 008 853

A paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, April, 1974.

Test Item Arrangement and Adaptation Level

Henry D. Olsen, Ph.D.
Medgar Evers College, CUNY

and

Robert S. Barickowski
Ohio University

The preponderance of literature indicates that the arrangement of items on a test according to item difficulty has no significant effect on test performance (Brenner, 1964; Flaugher, Melton and Myers, 1968; French and Greer, 1964; Marso, 1970; Monk and Stallings, 1970; Munz and Smouse, 1968 Smouse and Munz, 1968; Sax and Cromack, 1966). That is, an S would obtain the same score on a test with any one of the following item arrangements: 1) random order of item difficulty (R), 2) ascending level of difficulty, easy items followed by more difficult items (E-H), 3) descending level of difficulty, difficult items followed by easier items (H-E). Three studies (Hambleton, 1968; Lund, 1953; MacNicol, 1956) with experimental evidence, and several authors of measurement texts (Davis, 1951; Gronlund, 1965; Stanley and Ross, 1954) without referenced experimental evidence, recommend that test items be arranged in the easy to hard format. Although the preceding text book authors were partly refering to speeded tests, and other tests where item difficulty is used to discourage the Ss from continuing (e.g., Scholastic Aptitude Test, Graduate Record Examinations, etc.), their advice has been applied to all test forms. Finally, Heim (1955) found that Ss scored significantly higher on tests where items were arranged H-E, compared with E-H arrangement.

The purpose of the present study is to consider the problem of item arrangement in light of Helson's (1930) adaptation level theory. Studies involving adaptation level are generally concerned with supplying evidence which may help to answer the question: "Why do things appear as they do?" Helson (1964) indicates that, "Judgements are relative to prevailing norms or adaptation levels. Thus a 4-ounce fountain pen is heavy, but a baseball bat to be heavy must weigh over 40 ounces (p.26)." Current trends and issues in adaptation level theory cover a broad range from psychophysics to social psychology (Helson, 1964).

A researcher who applied adaptation level theory in a study of item arrangement would predict that Ss taking a test on which the item order was H-E would "adapt" to the hardest items so that easier items would subjectively seem even easier. With an E-H arrangement of test items the S would adapt to the easy items so that subsequent difficult items would appear to be more difficult than these same items in the H-E context. If perception of an item affects how one answers the item, the researcher would also be interested in the socres of Ss taking the H-M-E test compared with Ss taking these same items arranged E-M-H.

None of the studies concerned with the item-order effects considered the S's perception of the items he was attempting to answer. It is possible that, although items were arranged in order of difficulty, the Ss did not perceive the items as having different degrees of difficulty. That is, the Ss may not have perceived any difference between difficult items (e.g., median difficulty, PM about .20); medium items (e.g., PM about .50), and easy items (e.g., PM about .80). If this were the case, the expected effects predicted by adaption level theory would not have occured.

Furthermore, none of the preceeding item order studies took advantage of multivariate statistical techniques in analyzing the test results. Multivariate techniques could be applied to this kind of data by considering sets of easy items, medium difficulty items, and hard items as subtests (dependent variables) in a multivariate analysis of variance.

This study was designed to provide answers to the following questions:

1. Given a test on which items have been set in an easy, medium, and hard (E-M-H) arrangement, and a test having the same items arranged H-M-E, will Ss perceive the items on these test differentially? Specifically, will Ss who have the H-M-E arrangement perceive the H, M, and E sets of items as being significantly easier than Ss who have these same sets of items in the E-M-H context?

2. Does an S's perception of a set of items affect his score on the items?

3. Will the answers to question 1 and 2 be consistent across different item types? Will Ss have the same perceptions and score in the same manner when multiple choice or true-false items are used?

METHOD

The students from three sections of Teaching Reading and Language Arts in the Elementary School, Education 310, at Ohio University served as Ss for the study. The sections, taught by the senior author, consisted of 25,17, and 43 junior level students, for a total of 85Ss. The Ss had prepared for a midterm examination covering the basic components of reading instruction in the elementary grades (i.e., comprehension, word attack skills, material selection, individualization, diagnosis).

The items for the midterm examination were selected from a pool of 140 items given to 285 Ss during the preceding 1972-73 Fall quarter. True-false and multiple choice items were selected on the basis of their item difficulty and descrimination[2] indices. Previous investigators have found item difficulty (Brenner, 1964; Carter, 1942; Davis, 1951; Gibbons, 1940) and item descrimination (Brenner, 1964) values to be highly reliable. Table 1 presents the original item pool and the midterm examination medians of the item indices for the true-false and multiple choice items on the H, M, and E subtests.

Table 1 about here

Two forms of the examination were prepared[3]. The items on the H-M-E form were arranged as follows: 10 hard multiple choice items (H-MC), 10 hard true-false items (H-TF), 10 medium M-C items (M-MC), 10 medium T-F items (M-TF), 10 easy M-C items (E-MC), and 10 easy T-F items (E-TF). The items on the E-M-H form were arranged in reverse order of item difficulty, but in the same order of item type (i.e., multiple choice followed by true-false).

During the class session prior to the midterm examination the instructor (the junior author) told the Ss that following each item on the exam would be a Likert scale on which they were to rate each of the items. The scale consisted of the choices (1) very easy, (2) easy, (3) average, (4) difficult, and (5) very difficult. The Ss were told that if they conscientiously rated each item, the results would be helpful in retaining, deleting or adjusting each item for future examinations.

On the day of the midterm the instructor again reminded the Ss of the Likert scale and told them that the ratings would be of most value if they answered the questions in the order presented. During the exam the instructor, and a proctor, did not

observe anyone who was not complying with the directions.

The examinations were randomly distributed in each classroom, resulting in 42Ss taking the E-M-H arrangement and 43 Ss taking the H-M-E arrangement. There were 60 items and 60 ratings to be made, therefore, each S was asked to make 120 responses. The tests were electronically scored.

A multivariate analysis of variance was used to analyze the data. If this type of analysis yields significant results, univariate t-tests can be run on the subtest means (Hummel and Sligo, 1971). Therefore, the following procedures were used to gain a rough *a priori* estimate of the power of the statistical tests. A "medium" effect size (Cohen, 1969), a measure of the effect one desires to detect, of .50 was selected for this study. Cohen (1969, p.28) indicated that given = .05, n = 42 and an effect size of .50, that the power for a one-tailed independent t-test would be .74. That is, population mean differences of one-half standard deviation would be detected three out of four times in this study.

RESULTS AND DISCUSSION

Hotelling's T^2 (Morrison, 1967), the multivariate analogue of the univariate t-test, was used to analyze the data. The twelve dependent variables in the analyses consisted of the six parts H-MC, H-TF, M-MC, M-TF, E-MC, E-TF of the midterm, and rating scores arrived at by summing the ratings of the items in each part.

In the analysis the overall multivariate test was significant (tabled F (.05; 12, 72) 1.92; calculated F = 7.58) and, therefore, the univariate t-tests on each dependent variable were considered. Table 2 presents the means from each group, the pooled standard error of the mean difference s, and the univariate t-test for the six ratings.

Table 2 about here

The results in Table 2 indicate that the SS perceived the E, M, and H multiple choice items as being significantly easier when attempted in a H-M-E context then when these same items were attempted in an E-M-H context. When true-false items were considered, only the difficult, H, items were viewed as being significantly easier when viewed in an H-M-E context compared with the E-M-H context. Therefore, the first research question may be answered in the affirmative for multiple choice items and difficult true-false items.

The trend of the means in Table 2 suggests that with more E and M type true-false items or with larger sample size, significant differences might be found between the mean perceptions of the E and M type true-false items. That is, across all subtests the Ss perceived the items in the H-M-E context as being easier than items in the E-M-H context, but all of the differences were not significant.

Table 3 about here

Table 3 presents the means for each group, and the univariate t-tests for six subtests. The results presented in Tables 2 and 3 indicate that although the students preceived most of the items as being easier in the H-M-E context, there were no significant differences in the test scores on five of the six subtests. This result is in agreement with the preponderance of literature concerned with the topic of item arrangement. Only in the case of the E-MC subtest were the resultant means in the same direction as the perceived means. Since this result was not consistent with the results of the other subtest means, its support must be held in abeyance until further replication of this study can be made.

Further research in this area might be done on groups of Ss who have been differentiated on a pretest as having different levels of adaptation. Observation of the individual S data in this study indicates that some Ss may adapt "easily" to the item difficulties and some may not. For example, one S who took the E-M-H test had perceived scores on the E, M, H multiple choice subtests of 1.6, 3.0, 3.6 respectively; another S had scores on these same tests of 2.7, 2.7, 2.8. It might be conjectured that Ss who do adapt in the former manner, "easily", to item difficulties would score differently than Ss who do not. Munz and Smouse (1968) did find that interactions existed between personality variables and item arrangements.

This study should also be replicated across other populations of Ss and content areas. It may be that other Ss (e.g., elementary school children) will adapt to item arrangement, and that their scores will be affected.

FOOTNOTES

1 "Adaptation level" or "AL"- "the hypothesized neutral point or region of organic functioning at which stimuli coinciding with AL are indifferent or ineffective, stimuli above AL have a given character, and stimuli below AL have an opposite or complementary quality. AL represents the pooled affect of three classes of factors: (1) stimuli immediately responded to, or in focus of attention; (2) stimuli having background or contextual influence; and (3) residuals from past experience with similar stimuli" (English and English, 1958, p. 11.)

2 The index of discrimination was calculated using the net D method (Marshall and Hales, 1971, p. 230).

3 A better procedure would have been to use eight forms of the examination so that the true-false and multiple choice item sets would have been counter-balanced.

4 This procedure will yield only a rough estimate of power since the calculations should be based on the multivariate model. However, the authors know of no a priori means of selecting an effect size for this model.

Table 1

Original Item Pool and Midterm Item Indices on the Six Subtests of the Midterm Examination

Item Subtest		Median Difficulty		Median Discrimination	
		Item Pool	Midterm	Item Pool	Midterm
Multiple Choice	Easy	.85	.76	.14	.17
	Medium	.59	.50	.23	.26
	Hard	.19	.17	.15	.14
True-False	Easy	.96	.94	.06	.04
	Medium	.59	.76	.28	.22
	Hard	.31	.36	.15	.20

Table 2

Group Means, Pooled Standard Error of the Mean Differences and Univariate t-Tests for the Item Ratings

Item Subtest		Subtest Mean**		Pooled Standard Error of the Mean Differences	t
		H-M-E	E-M-H		
Multiple Choice	Easy	2.81	3.02	.10	2.18*
	Medium	3.03	3.29	.10	2.59*
	Hard	3.07	3.70	.10	6.51*
True-False	Easy	2.28	2.46	.11	1.42
	Medium	3.04	3.19	.10	1.61
	Hard	2.91	3.39	.11	4.49*

** 1 = very easy, 5 = very difficult,

*Significant at α = .05; $t(.05; 83) < 2.00$.

Table 3

Group Means, Pooled Standard Error of the Mean Differences and Univariate t-Tests for the Subtests

Item Subtest		Subtest Mean		Pooled Standard Error of the Mean Differences	t
		H-M-F	E-M-H		
Multiple Choice	Easy	7.72	6.19	.32	4.85*
	Medium	4.93	5.36	.35	1.20
	Hard	2.13	1.93	.30	.79
True-False	Easy	9.63	9.43	.14	1.39
	Medium	6.91	7.12	.30	.71
	Hard	3.70	3.74	.35	.12

*Significant at $\alpha = .05$; $t(.05; 83) < 2.00$.

REFERENCES

BRENNER, M.H. Test difficulty, reliability, and discrimination as functions of item difficulty order. Journal of Applied Psychology, 1964, 48(2), 98-100.

CARTER, H. How reliable are the common measures of difficulty and validity of objective test items. Jounral of Applied Psychology, 1942, 13, 31-39.

COHEN, J. Statistical power analyses for the behavioral sciences. New York: Academic Press, 1969.

DAVIS, F.G. Item selection techniques. In E.F. Lindquist (Ed.) Educational Measurement. Washington, D.C.: Council on Education, 1951, 266-328.

ENGLISH, H.B., & ENGLISH, A.C. A comprehensive dictionary of psychological and psychoanalytical terms. New York: David McKay, 1958.

FLAUGHER, R. L., MELTON, R. S., & MEYERS, C. T. Item rearrangement under typical test conditions. Educational and Psychological Measurement, 1968, 28, 818-824.

FRENCH, J.L., & GREER, D. Effect of test-item arrangement on physiological and psychological behavior in primary-school children. Journal of Educational Measurement, 1964, 1(2), 151-53.

GIBBONS, C.C. The predictive value of the most valid items of an examination. Journal of Educational Psychology, 1940, 31, 616-621.

GRONLUND, N.E. Measurement and evaluation in teaching. New York: MacMillan, 1965.

HAMBLETON, R.E. The effect of item order and anxiety on test performance and stress. ERIC, 1968, 017 960.

HEIM, A.W. Adaptation ot level of difficulty in intelligence testing. British Journal of Psychology, 1955, 46, 211-224.

HELSON, H. The nature and problem of perception. In Wheeler, R.H. (Ed.) Readings in Psychology. New York: Crowell, 1930, Chapter 20.

HUMMEL, T.J., & SLIGO, J.R. Empirical comparison of univariate and multivariate analysis of variance procedures, Psychological Bulletin, 1971, 76(1), 49-57.

LUND, K. Test performance as related to order of item difficulty, anxiety, and intelligence. Unpublished doctoral dissertation, Northwestern University, 1953.

MACNICOL, K. Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, Educational Testing Service, 1956.

MARSHALL, J.C., & HALES, L.W. Classroom test construction. Reading, Mass.: Addison-Wesley. 1971

MARSO, R.N., Test item arrangement, testing time, and performance. Journal of Educational Measurement, 1970, 7(2), 113-118.

MONK, J.J., & STALLINGS, W.M. Effects of item order on test scores. Journal of Educational Research, 1970, 63(10), 463-464.

MORRISON, D.F. Multivariate statistical methods. New York: McGraw-Hill, 1967.

MUNZ, D.C., & SMOUSE, A.D. Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. Journal of Educational Psychology, 1968, 59(5), 370-374.

SAX, G., & CROMACK, T.R. The effects of various forms of item arrangements on test performance. Journal of Educational Measurement, 1966, 3(4), 309-311.

SMOUSE, A.D., & ,UNZ, D.C. The effects of anxiety and item difficulty sequence on achievement testing scores. Journal of Psychology, 1968, 68, 181-184.

STANLEY, J.C., & ROSS, C.C. Measurement in today's schools. New York: Prentice-Hall, 1954, 139-162.